A Predictive Model for Advertiser Value-Per-Click in Sponsored Search

Eric Sodomka*
Brown University
Providence, Rhode Island

Sébastien Lahaie*
Microsoft Research
New York, New York

Dustin Hillard*
Microsoft Corp.
Redmond, Washington

ABSTRACT

Sponsored search is a form of online advertising where advertisers bid for placement next to search engine results for specific keywords. As search engines compete for the growing share of online ad spend, it becomes important for them to understand what keywords advertisers value most, and what characteristics of keywords drive value. In this paper we propose an approach to keyword value prediction that draws on advertiser bidding behavior across the terms and campaigns in an account. We provide original insights into the structure of sponsored search accounts that motivate the use of a hierarchical modeling strategy. We propose an economically meaningful loss function which allows us to implicitly fit a linear model for values given observables such as bids and click-through rates. The model draws on demographic and textual features of keywords and takes advantage of the hierarchical structure of sponsored search accounts. Its predictive quality is evaluated on several high-revenue and high-exposure advertising accounts on a major search engine. Besides the general evaluation of advertiser welfare, our approach has potential applications to keyword and bid suggestion.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Economics; K.1 [**The Computer Industry**]: Markets

Keywords

Sponsored search; Hierarchical model; Regret loss function

1. INTRODUCTION

Online advertising spend exceeded $100 billion for the first time in 2012, with a significant fraction going to advertising on search engines, a segment known as sponsored search.[1] Sponsored search refers to the practice of displaying ads alongside search results whenever a user issues a query. Advertisers develop campaigns by selecting the keywords they wish to advertise on and setting bids for those keywords, with the placement and cost of ads determined via an auction process [15]. Developing a high-performance online ad campaign is a complicated task. Advertisers must select keywords and bids to optimize returns, and it may require costly experimentation to uncover the keywords that yield the highest profits per click.

This paper proposes a hierarchical linear model to infer an advertiser's value per click on search terms. Our choice of model is based on the hypothesis that keyword values are linked to characteristics such as the demographic profile of users they attract. The growing trend towards geographic, demographic, and even behavioral targeting is evidence that these characteristics can be closely linked to returns [21].

Our model is composed of two key features. First, the regression model uses an *economically meaningful* loss function to determine the regression coefficients. Keyword auctions are not truthful, so bids cannot be taken as proxies for values—independent work has shown that values can be as high as 125% of bids [7]. However, while an advertiser's actual value per click is unobservable, it is possible to fit a model of values such that the advertiser's observed bidding behavior, under the values imputed by the model, is as close to rational as possible in the sense that it minimizes foregone profit, or *regret*. We explain that our regret-based loss function falls in the well-known class of Bregman divergences used in machine learning [2], and therefore allows for efficient convex optimization algorithms to fit the model. To define the loss function on each keyword we need a model of advertiser beliefs about the clicks and costs that are obtained at different bids. For this we draw on the recent work of Pin and Key [16] to develop click and cost function estimates, and in doing so provide an independent evaluation of their approach.

Second, our model exploits the *hierarchical structure* of advertiser accounts to better estimate keyword values. Advertiser accounts consist of various levels of organization, with different decisions (such as demographic targeting or bidding) being made at each level. The advertiser account structure introduces commonality in keyword characteristics, and therefore likely indicates commonality in keyword values. We show that advertiser account structures demonstrate the kind of skew in campaign sizes and bids that motivate hierarchical modeling, and that our hierarchical regression model can improve predictions over a baseline that only takes into account local account structure.

Information about bidder values can inform many aspects of the keyword auction design, including changes to the ranking rules to improve revenue and reserve pricing policies [14]. We see two immediate applications of the value estimates provided by our approach: keyword and bid sugges-

*This work was done while the authors were at Yahoo! Inc.
[1] `www.emarketer.com/Article.aspx?R=1009592`

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

tion. Search engines typically provide keyword suggestion tools to help advertisers augment their campaigns. The current state of the art provides keyword suggestions based on statistical and semantic similarities using a campaign's initial set of keywords [6], but we have not found any research on how to filter and rank keyword suggestions according to value to the advertiser. The value estimates from our model provide a principled ranking criterion.

The only work on bid generation we are aware of is the recent paper by Broder et al. [5], who use machine learning to directly predict bids based purely on textual features of keywords and ads. While they report good prediction performance, their approach to recommending bids cannot react to changes in opponent bids, and does not provide a criterion for ranking suggestions—high bids may indicate the most competitive keywords, rather than the most valuable. By uncovering the primitives behind advertiser behavior (i.e., values) it becomes possible to automate the complete process of keyword ranking and bidding.

To summarize, our work makes the following four contributions, with the evaluation of our approach being the main contribution.

- Original insights on the structure of sponsored search accounts that motivate hierarchical regression modeling (Section 3).
- A regret-based modeling strategy that allows one to fit a model of unobserved values based on observed bids and click-through rates (Section 4).
- An independent evaluation of the Pin and Key [16] approach to modeling advertiser beliefs on click and cost functions (Section 5).
- An experimental evaluation of our modeling approach using real sponsored search account data (Section 6).

The remainder of this section reviews related work. Section 2 provides the background on sponsored search needed to follow the paper, while Section 7 concludes with directions for improvement and future work.

Related work. Regression models of account performance (e.g., click-through rates) have appeared in the marketing literature for single accounts [11, 13]. Building such models on the search engine side can be an insightful exercise because, while the search engine may not have conversion data, it may have much finer-grained information about the user traffic that visits the ads. The paper of Rutz and Bucklin [17] is most closely related to ours in that it expressly addresses the problem of estimating values, in their case as implied by conversions. Using data from the paid search campaign of a hotel chain, they apply several logit models to predict conversions based on features such as the presence of brand or geographic information.

Our work also connects with the small but influential economic literature on equilibrium models of sponsored search. Edelman et al. [9] introduced the solution concept of envy-free equilibrium, while Varian [20] showed how it could be applied to derive bounds on values per click using bid data from Google. Athey and Nekipelov [1] develop a model that incorporates uncertainty in competitors and quality scores in order to provide more refined bounds and even points estimates. Pin and Key [16] develop a similar method that is much more scalable, and we draw on their work to estimate clicks and cost as functions of bid.

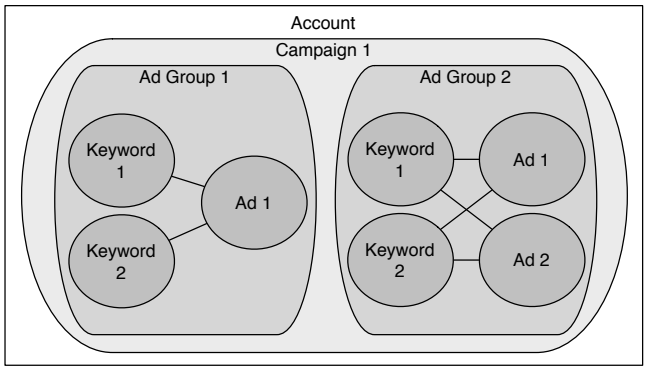


Figure 1: Hierarchical structure of a sponsored search account.

2. SPONSORED SEARCH

We now describe the process of sponsored search and the associated terminology used in this paper. At a high level, the process of bidding in sponsored search proceeds as follows: each advertiser specifies a list of **keywords** (or **terms**), a standing **bid** for each of those keywords, and a specific advertisement (or **creative**) they wish to display for each keyword. When a user issues a query to the search engine, an auction is run among the relevant advertisers to determine which ads appear, in what order on the page they appear, and how much money each advertiser must pay.

The ranking of an ad is determined by a combination of its bid and **quality score**, which is meant to capture the ad's relevance to the keyword; an important ingredient in this score is the search engine's estimate of the ad's probability of being clicked, known as the **click-through rate (CTR)**. By convention, the payment scheme is *per click*, meaning the advertiser only pays when its ads are clicked, not simply when they are shown; the price of a click is often called the **cost per click (CPC)**. Advertisers receive some expected **value per click**, which is private information, and together with the advertiser's estimates of CTR and CPC, drives their decision making.

Advertisers can also make additional decisions beyond what to bid to obtain finer control over who sees their ads. One popular refinement is the choice of **standard** or **advanced** match. For example, if advanced match is enabled for the keyword "sports shoes", the search engine can show the same ad for the related queries "running shoes" or "track shoes" rather than just that exact query. Advertisers can also impose **targeting** settings to specify that their ads should only be shown to users from specific locales or demographics. For example, an advertiser might target their ad to only appear to male searchers in their thirties from California. Finally, advertisers can set **budgets** to limit the amount they spend in a given day; when the advertiser hits its budget, it drops out of the day's auctions. The stochasticity these additional settings create across auctions can be leveraged to obtain models of advertiser beliefs over clicks and costs [1, 16].

Search engines provide advertisers with some organizational structure for their accounts to enable them to apply targeting and budgeting decisions to several keywords at once. Figure 1 illustrates the hierarchical structure of a sponsored search account, which is shared among the leading search engines. Keywords and creatives can are grouped

together into **ad groups**; all keywords in the ad group display the same rotation of ads. Advertisers are also able to place a single bid at the ad group level, and all keywords with unspecified bids will default to that ad group bid. Ad groups are grouped into **campaigns**; budgets and targeting options are typically set at the campaign level. The hierarchical structure of these accounts should provide information about an advertiser's keyword values, since keywords in the same ad group or campaign share features as defined by the account hierarchy. This account structure provides partial motivation for hierarchical regression modeling.

3. PRELIMINARY ANALYSIS

Our data set consists of Yahoo's sponsored search logs over one month in the first half of 2010. For each query in the data set, we have information about the *auction*, *displayed advertisers*, and *user* (who issued the query). Auction information includes the query and the number of ads displayed at the top and to the right of the page. Advertiser information includes each advertiser's bid, whether the ad was displayed via exact or advanced match, the original keyword each advertiser bid on, the displayed creatives, and the predicted CTRs decomposed into position and advertiser effects. User information includes demographics such as predicted age, gender and zip code.

We focus our investigation on 100 advertiser accounts sampled from the set of all accounts. We obtained the total impressions, clicks, and revenue (for Yahoo, or equivalently, advertiser costs) for each account in Yahoo's database and found that, across accounts, impressions and revenue follow lognormal distributions while clicks follow a Pareto distribution (i.e., power law). In particular, the top 10% of accounts by click volume are responsible for over 80% of the monthly clicks, with a similar skew in the distributions for impressions and revenue. Therefore, uniform sampling is inappropriate because the sample would be overwhelmed by accounts with low click volume and revenue.

We found that click volume has a strong (linear) correlation of at least 0.5 with impression volume and revenue. Therefore, we choose to sample accounts proportional to click volume so as to strike a balance between high-revenue and high-exposure accounts. Note that these accounts are not intended to be a representative sample of the entire account space. Rather, they are meant to be a representative sample of the accounts for which it is most worthwhile to provide value estimation services: their high revenue makes them valuable to the search engine, and their high click volume indicates their ads are relevant to users. Our 100 sampled accounts are responsible for hundreds of thousands of clicks throughout the month, and taken together they contain nearly 150K terms. The remainder of this section provides a detailed examination of the structure of the accounts in our sample, which serves to motivate and inform our hierarchical modeling strategy.

3.1 Account Structure

A summary of the basic structure of the sampled accounts is given in Table 1. First, note that even though we sampled proportional to clicks, some accounts are very small: just three terms or one ad group. Further filtering of accounts is needed to restrict our attention to accounts with enough terms to model. An important observation is that the median and mean number of terms per ad group is very small; this makes sense considering the terms in an ad group share the same ads. This suggests that ad groups are very homogeneous, and we would expect clicks from different terms in an ad group to have similar values to the advertiser. Once one moves to the campaign level the number of terms starts to be large enough to support model fitting.

metric	min	median	mean	max
terms per ad group	1	2	6.3	100
terms per campaign	1	30	145.4	4000
ad groups per campaign	1	10	30.9	400
terms per account	3	400	1456.0	10000
ad groups per account	1	300	431.0	8000
campaigns per account	1	20	25.5	100

Table 1: Summary of account structure across the 100 sampled accounts. The min, median, and max have been rounded to one significant figure.

A hierarchical structure to the data alone does not completely motivate hierarchical modeling. The latter is useful when groups have uneven sample sizes and some groups are small. In that case the group intercepts are pulled towards those of larger groups, so that information from large groups is taken into account for predictions within small groups, which otherwise do not have enough data to support inference.

Figure 2 provides plots that describe the distributions of ad group and campaign sizes across the accounts. Recall that skewness is a measure of the asymmetry of a distribution. We see that ad group sizes have a positive skewness for almost all accounts, which indicates a long right tail—there are a few large ad groups and many small ones, rather than the reverse. The kurtosis is a measure of the sharpness of a distribution's peak. Most accounts have ad group sizes whose kurtosis exceeds that of the normal distribution, which means accounts consist mainly of relatively small ad groups. The same pattern holds for campaign sizes, which show a positive skewness and high kurtosis in general, though not to the extent of ad group sizes. This justifies a hierarchical model that includes every level of an account.

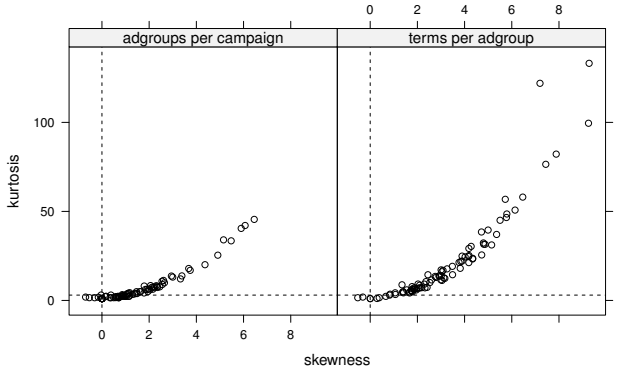


Figure 2: Skewness and kurtosis of group (ad group and campaign) sizes for the 100 accounts. The reference lines show the standard moments of the normal distribution: 0 for skewness and 3 for kurtosis.

3.2 Tail Contribution

Our observations on account structure might lead one to believe that only a few large ad groups matter in an account, but this is incorrect. In aggregate, small ad groups typically make up a substantial fraction of an account's click volume and revenue, so that it remains important to model advertiser value on terms in small ad groups. To confirm this, we examine the tail of the distribution of impressions, clicks, and revenue across the ad groups in an account, where the tail is defined as the bottom 80% of ad groups for the associated metric. We define three regimes: the 'Pareto' regime corresponds to a tail contribution of less than 20% to the total; the 'Long-Tail' regime corresponds to a contribution of at least 50%; and 'Intermediate' lies between the two.

We find that around 30 accounts fall under the Pareto regime for clicks or for revenue; thus, we cannot focus on just the head (top 20%) of ad groups in general. There are 13 accounts that fall under the Long-Tail regime for both clicks and revenue, and 19 accounts that are Long-Tail for at least one of them. Furthermore, the accounts in the Pareto regime are not responsible for much click volume or revenue in our account sample (less than 15%). We conclude that tail ad groups and terms, in aggregate, often generate much of the click volume and revenue in an account, and it is therefore important to develop models of value on all the terms and ad groups.

3.3 Bid Variation

We now examine the bid variation in accounts. We first consider bid changes across time. For each term we counted the number of distinct bids placed on that term throughout the month, and averaged over terms in an account. The numbers are low: the maximum number of distinct bids is 60, or just 2 new bids per day.[2] The relative standard deviation (standard deviation over the mean, expressed in percents) of the bids on a term, averaged over the terms in an account, is also very small: the mean is 3% over the accounts. For these reasons we choose to consider just the average bid across time on a term—a similar simplification was made by Broder et al. [5], who developed regression models of bids in sponsored search.

We next consider bid variation across the terms in an account; we would like to understand at what level (ad group, campaign, account) the variation arises. To decompose the variance among levels we make use of a simple hierarchical model that also serves as a precursor to our later model for values. Let b_i be the bid on term i. We model the bids in ad group j as drawn from a normal distribution with mean b_j. The means b_j of the ad groups in a campaign k are themselves drawn from a normal distribution with mean b_k, and the campaign means are normal with mean b_h:

$$b_i \sim \mathcal{N}(b_j, \sigma_1^2), \quad b_j \sim \mathcal{N}(b_k, \sigma_2^2), \quad b_k \sim \mathcal{N}(b_h, \sigma_3^2).$$

Here it is implied that term i is in ad group j, which is in campaign k. The proportion of variance at the first (i.e., ad group) level is $\sigma_1^2/(\sigma_1^2 + \sigma_2^2 + \sigma_3^2)$, and similarly for the other levels. For each account, the hierarchical model was fit using the `lmer` function in R. The assumption of normal distributions amounts to fitting a model with squared loss. Figure 3 gives the variance proportions for the sampled accounts. We find a wide spectrum of proportions: for some accounts almost all the variance occurs at the ad group level, while for others it is all at the campaign or account level.

[2]This does not necessarily mean that no interesting bidding behavior occurs across time—advertisers could be using intricate strategies that cycle through different bids. This has been observed in some early studies [8].

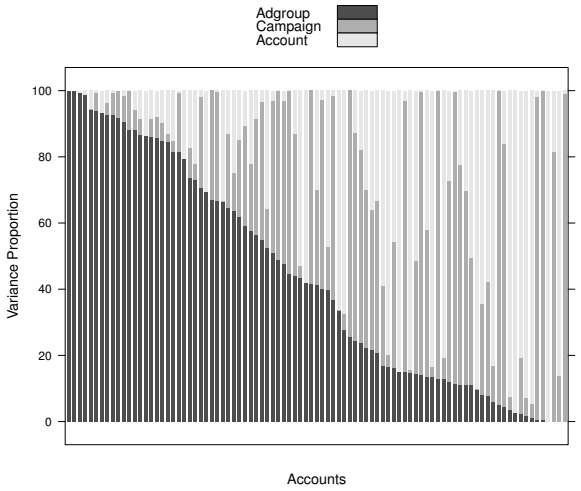


Figure 3: Proportion of bid variance that arises among terms within ad groups, ad groups within campaigns, and campaigns within accounts.

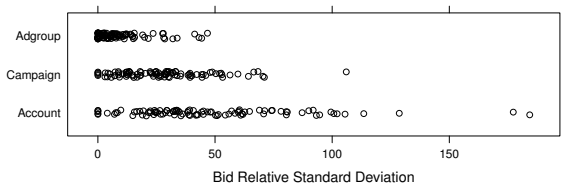


Figure 4: Relative standard deviation of bids. For each account, we take the relative standard deviation of bids within its ad groups, then average these over the ad groups in the account. We do the same for campaigns. The final line lists the relative standard deviation of bids over terms in each account.

The amount of bid variation at each level is also informative. For each account we looked at the relative standard deviation of bids within each ad group, and averaged over ad groups. We did the same looking at bids within campaigns, and within the account. The results are given in Figure 4. We find that relative standard deviation within ad groups is small: the median across accounts is 5%. The median within-campaign and within-account relative standard deviations are 26% and 41% respectively. This seems to suggest that many advertisers make their bidding decisions at the ad group level, and the bids on terms within an ad group do not stray far from their ad group baseline. The most interesting and challenging inference problem in such cases is to predict ad group-level average values rather than term-specific values.

4. HIERARCHICAL MODEL

The basis for our value estimation approach is a simple quasi-linear model of advertiser utility on each term. We focus on an individual advertiser with n terms in its account, with i used to index terms. Let v_i be the advertiser's value per click on term i. Let $c_i : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ be an extended real-value function that gives the expected cost per impression $c_i(x_i)$ of obtaining a click-through rate (CTR) x_i on term i. The purpose of introducing $+\infty$ into the range is to implicitly encode the domain of c_i as $\{x_i \in \mathbf{R} : c_i(x_i) < +\infty\}$, following standard conventions in convex analysis. For instance, negative CTRs are infeasible and would have a cost of $+\infty$. We assume that c_i is convex and differentiable.[3] The cost functions estimated later in Section 5 are in fact piece-wise linear, but by selecting a subgradient at the (finite) number of points of non-differentiability, the following applies with minor technical changes.

We assume that the advertiser's utility for clicks on a term is quasi-linear in cost, so that it takes the form

$$u_i(x_i, c_i) = v_i x_i - c_i(x_i). \tag{1}$$

Letting w_i be the search volume for term i, the aggregate utility to the advertiser of obtaining the vector of CTRs $x = (x_1, \ldots, x_n)$ across the terms in its account, given the vector of cost functions $c = (c_1, \ldots, c_n)$, is $u(x, c) = \sum_{i=1}^{n} w_i u_i(x_i, c_i)$. Our goal is to obtain a regression model of the v_i values, but these are not observable. Instead, we observe the CTRs x_i chosen by the advertiser via its bids on each term, and we can estimate the cost functions c_i.

4.1 Loss Function

A naive approach to developing a regression model of values would be to first estimate the values v_i on each term, and then run a standard regression on top of these (e.g., using least squares). Indeed, from the utility form (1) and our convexity assumptions on c_i, it follows that at the advertiser's observed choice of CTR x_i we must have $v_i = \nabla c_i(x_i)$ under utility-maximizing behavior (where ∇ refers to the first derivative).

The issue with this approach is that typical loss functions like squared error find little justification in economic settings like sponsored search: variance in the estimated values within an ad group may be a result of optimization error on the part of the advertiser, rather than statistical error, and there is no sound basis for optimization errors to be Gaussian [19]. Here we develop an economically meaningful loss function that is also computationally appealing because the problem of model fitting remains convex, as with conventional statistical loss functions. Under our loss function, value estimation occurs in tandem with model fitting, rather than as a preliminary step.

The idea is to draw on the notion of Bregman divergence, which has seen increasing attention in machine learning [2]. First, we need the concept of a convex conjugate. In what follows, we will suppress the term subscript i on values, costs, and CTRs for clarity. The *dual space* $\mathbf{R}^*$ is the vector space of all linear functions on $\mathbf{R}$. Observe that values can be identified with elements of the dual space. The *convex conjugate* $c^* : \mathbf{R}^* \to \mathbf{R}$ of cost function c is defined as

$$c^*(v) = \sup_{x \in \mathbf{R}} \{vx - c(x)\}. \tag{2}$$

Observe that c^* is convex, even if c is not, because is it the supremum of affine functions. A Bregman divergence is defined with respect to a strictly convex, differentiable function, in our case the conjugate cost function c^*. Given c^*, the divergence between two values v and v' in its domain is defined as

$$D_{c^*}(v'||v) = c^*(v') - c^*(v) - \nabla c^*(v) \cdot (v' - v). \tag{3}$$

This is the loss function between values that we propose. Here v' should be viewed as the estimated value, and v the true value (not directly observed).

The following proposition collects some standard facts about Bregman divergence (e.g., see [2]) that serve to motivate this choice of loss and explain how it is evaluated in practice.

PROPOSITION 1. *Let c^* be a convex and differentiable function. The associated Bregman divergence D_{c^*} satisfies the following properties.*

1. *$D_{c^*}(v'||v) \geq 0$ for all v, v' in the domain of c^*, with equality if $v = v'$.*
2. *$D_{c^*}(v'||v)$ is convex in its first argument.*
3. *$D_{c^*}(v'||v) = D_c(x||x')$ where x, x' are such that $v = \nabla c(x)$ and $v' = \nabla c(x')$.*

Property 1 confirms that D_{c^*} is a sensible loss function. Bregman divergence in fact generalizes losses such as squared loss, KL-divergence, and Itakura-Saito distance, which can all be recovered with a suitable choice of convex function c [2]; for example, squared loss corresponds to a quadratic cost function. Property 2 ensures that is it computationally tractable to fit the estimate v'. Finally, Property 3 shows that our loss function can be equivalently viewed as a loss on CTRs, which are observable.

Economic Interpretation

To see the economic motivation for this loss function, let $v' = \nabla c(x')$ be the value for which the choice of CTR x' is optimal. Then by Property 3 loss function (3) evaluates to

$$\begin{aligned} D_c(x||x') &= c(x) - c(x') - \nabla c(x') \cdot (x - x') && (4) \\ &= [v'x' - c(x')] - [v'x - c(x)]. && (5) \end{aligned}$$

This is the *regret* from bidding so that CTR x is received, rather than the optimal choice x' when the advertiser's value is v'. Stated another way, in order to minimize the loss we seek an estimated value v' such that the advertiser's regret from obtaining the observed CTR x under this value is minimized.

Varian [19] has proposed several 'economically meaningful' loss functions of this sort for both parametric and non-parametric models. For parametric models of production analysis, he proposes a loss function which captures the degree to which the observed choice behavior fails to maximize the *estimated* production function, which is precisely the re-

[3] Since agent values are linear in CTR, replacing c_i with its convex envelope (the largest convex function it dominates) does not change the agent's utility maximization problem. Therefore, the assumption that c_i is convex is without loss of generality when it comes to analyzing an agent's choice of bid and CTR on a keyword.

gret in (5).[4] In microeconomics the conjugate (2) is known as the *indirect utility function*; intuitively, it gives the maximum utility that an advertiser with value v can achieve. Our loss function is the Bregman divergence associated with indirect utility.

Evaluating Loss

The loss function (3) is defined in terms of unobservable values, whereas its alternate form (4) is defined in terms of the observable CTR and cost function, but does not allow one to incorporate a linear model for value. Instead we will work with the following form for loss. Let x be the optimal choice of clicks when the advertiser's value is v; from (2) and Danskin's theorem [3, p. 717] we have $x = \nabla c^*(v)$. Thus,

$$\begin{aligned} D_{c^*}(v'||v) &= c^*(v') - c^*(v) - \nabla c^*(v) \cdot (v' - v) \\ &= c^*(v') - [vx - c(x)] - x \cdot (v' - v) \\ &= c^*(v') - xv' + c(x). \end{aligned}$$

Now, if we have a linear model for the fitted value, so that $v' = a \cdot \beta$ where a is the feature vector and β are the coefficients, then the loss function can be written as

$$D_{c^*}(v'||v) = c^*(a \cdot \beta) - x(a \cdot \beta) + c(x). \tag{6}$$

This loss is convex in the parameters β and is formulated in terms of the observed chosen CTR x, the cost function c, and its conjugate c^*. The conjugate c^* of a one-dimensional convex cost function can be computed in linear time using elementary convex hull algorithms. Section 5 addresses the problem of estimating the cost function c.

4.2 Levels

We now describe the aggregate loss function across all terms in an account. There are three levels in an account: (1) terms within ad groups, (2) ad groups within campaigns, and (3) campaigns within the account. Our regret-based loss function applies to the first level, while standard squared loss is used at higher levels for regularization. We index the terms, ad groups, and campaigns in the advertiser's account by i, j, and k respectively, and with an abuse of notation write $i \in j$ to denote that term i belongs to ad group j, and $j \in k$ to denote that ad group j belongs to campaign k.

We implicitly observe v_i for each term i via the chosen CTRs x_i. Let a_i be the vector of predictors for term i and let v_j be the intercept for ad group j. Let β_1 be the model coefficients at the ad group level. Writing D_i for the divergence (3) associated with c_i^*, the aggregate loss at the first level is

$$\sum_k \sum_{j \in k} \sum_{i \in j} D_i(v_j + a_i \cdot \beta_1 || v_i). \tag{7}$$

As explained above each term in the summation can be evaluated in practice via (6), and because D_i is convex in its first argument it is also convex in the parameters β_1 and v_j.

Analogous aggregate loss functions could be derived for the second and third levels, but it is unclear whether advertisers reason in terms of regret at those levels. Another problematic aspect of using regret-based loss at higher levels is that parameters appear in the second argument, where convexity is not guaranteed. Because the purpose of higher-level loss is regularization, we therefore use standard squared loss. Again, let a_j be the vector of predictors for ad group j and let v_k be the intercept for campaign k. Let β_2 be the model coefficients at the campaign level. The loss at this level is

$$\sum_k \sum_{j \in k} (v_k + a_j \cdot \beta_2 - v_j)^2. \tag{8}$$

In the ad group-level loss (8), the "observation" v_j is in fact the intercept fit at the lower level in (7).

We use the index h to refer to the intercept at the third level (i.e., account level), where we also use squared loss:

$$\sum_k (v_h + a_k \cdot \beta_3 - v_k)^2. \tag{9}$$

The "observation" v_k here in fact corresponds to the intercept from (8). Finally, the aggregate losses from the three levels are summed up to yield the complete loss used to ultimately fit the model. Again we stress that only the aggregate regret (7) captures the performance of the model; losses at higher levels of the hierarchy are for regularization, as motivated by our observations on account structure in Section 3. Introducing (8) pulls all the v_j ad group intercepts towards a common intercept v_k for the campaign, so that information is shared across ad groups in the campaign, and similarly for (9). This is completely analogous to the motivation for multi-level models with standard losses like least-squares [10].

The aggregate loss combined over all levels is convex in all model parameters, so fitting the model is straightforward via algorithms like stochastic gradient descent. We implemented an incremental subgradient algorithm [3, p. 614] similar to the stochastic gradient descent used in [5] to develop a model of bids. While the implementation details are beyond the scope of this paper, we note that running times to fit the model ranged from under a second to around 25 minutes on the largest account (we did not attempt to optimize the code). The bottleneck is in reading in the account data and setting up the data structures rather than running the algorithm.

5. COST FUNCTION ESTIMATION

The preceding sections developed a linear model for estimating a term i's value-per-click from its features. The proposed method required knowledge of advertiser costs $c_i(x_i)$ for a given observed (implicit) choice of click-through rate x_i; we now describe an approach to estimating this cost function $c_i(x_i)$.

As discussed in Section 2, there are numerous factors that cause variability in the competitive landscape across auctions. These factors make it impossible to compute exact costs for the next upcoming search, making cost functions based on individual auctions (as studied by Edelman et al. [9] and Varian [20]) less realistic for our domain. The observation from Section 3 that advertisers infrequently modify their bids further suggests that advertisers make decisions at a coarser level of granularity than the auction level. Following recent approaches [1, 16], we assume the advertiser creates a stochastic model of costs based on a distribution over the observed competitive landscape.

[4] Our model of advertiser utility is closer to producer theory than consumer theory. We have surveyed the econometric literature and, despite Varian's convincing arguments, it seems the idea of 'economic' loss functions has not caught on, and standards like squared loss are still in favor. This may be due to computational reasons: not all of Varian's loss functions are convex in the estimated parameters.

Specifically, we use a version of the cost function described by Pin and Key [16].[5] The model is from the perspective of a single advertiser, with each term in the advertiser's account treated independently. Each opponent participating in an auction is assumed to stochastically submit a bid i.i.d. from a known probability density function; let $F(b)$ be the corresponding c.d.f. that gives the probability an opponent submits a bid less than b. These bids are assumed to be weighted to adjust for differences in opponent quality score[6]. Variation in $F(b)$ comes from both the variation in these quality scores across searches and from the search-specific set of competing advertisers (determined by standard match, advanced match and targeting settings described in Section 2).

Let k be a possible slot, where $k = 0$ is the top slot. Given the cumulative distribution over opponent bids $F(b)$, an advertiser can compute the probability ϕ of appearing in the kth slot in a single auction when there are n opponents and the advertiser places a bid b:

$$\phi(k, n, b) = \binom{n}{k}(1 - F(b))^k F(b)^{n-k} \tag{10}$$

In words, this computation is the number of ways k opponents could be chosen to appear in slots above the advertiser, times the probability that those k opponents appear in slots above the advertiser, times the probability the remaining $n - k$ opponents appear in slots below the advertiser.

For a given term, let p_n be the probability that n opponents participate in each auction, where n is at most N. Each slot k has a known click-through rate s_k, where $s_k = 0$ if k exceeds the number of available slots. Assuming the advertiser bids above the known reserve price r, the advertiser's expected click-through rate $\hat{x}_i(b)$ as a function of its bid b is computed as follows:

$$\hat{x}_i(b) = \sum_{n=0}^{N}\sum_{k=0}^{n} p_n \phi(k, n, b) s_k \tag{11}$$

Similarly, cost per impression $\hat{c}_i(b)$ is computed as:

$$\hat{c}_i(b) = \sum_{n=0}^{N}\sum_{k=0}^{n} p_n \phi(k, n, b) s_k \underbrace{\left(b - \int_r^b \frac{F(t)^{n-k}}{F(b)^{n-k}} dt\right)} \tag{12}$$

The additional bracketed term is the expected cost per click, given the advertiser appears in the kth slot against n bidders with a bid of b and reserve price r.

A given bid b is thus associated with an expected cost per impression $\hat{c}_i(b)$ and an expected click-through rate $\hat{x}_i(b)$. The expected cost per impression $c_i(x_i)$ as a function of click-through rate x_i is then reconstructed by pairing the calculated $\hat{x}_i(b)$ and $\hat{c}_i(b)$ values for each possible bid b:

$$c_i(x_i) = \hat{c}_i(\hat{x}_i^{-1}(x_i)) \tag{13}$$

[5]Pin and Key present several models of increasing complexity; we use their most complex model that does not consider distributions over promoted reserve price (i.e., reserve price for ads shown at the top of the page), since this data was unavailable.

[6]Recall from Section 2 that advertisers are ranked by their bid b times their quality score w. The opponent bid distribution $F(b)$ created by advertiser a gives the probability that advertiser a's weighted bid wb is greater than the opponent's weighted bid $w'b'$.

This version of the Pin and Key model matches our environment closely but not exactly. First, reserve prices are not known exactly and are not necessarily of a fixed value in our domain. Second, the slot click-through rates depend not only on the slot k, but also the number of promoted slots (i.e., shown at the top) and the number of opponents. To extend the model, we take additional summations in Equations (11) and (12) over possible reserve prices r and possible available promoted slots d, and slot click-through rate s_k is instead computed as s_{kdn}, which gives a different slot click-through rate depending on the competitive landscape.

5.1 Distribution Estimation

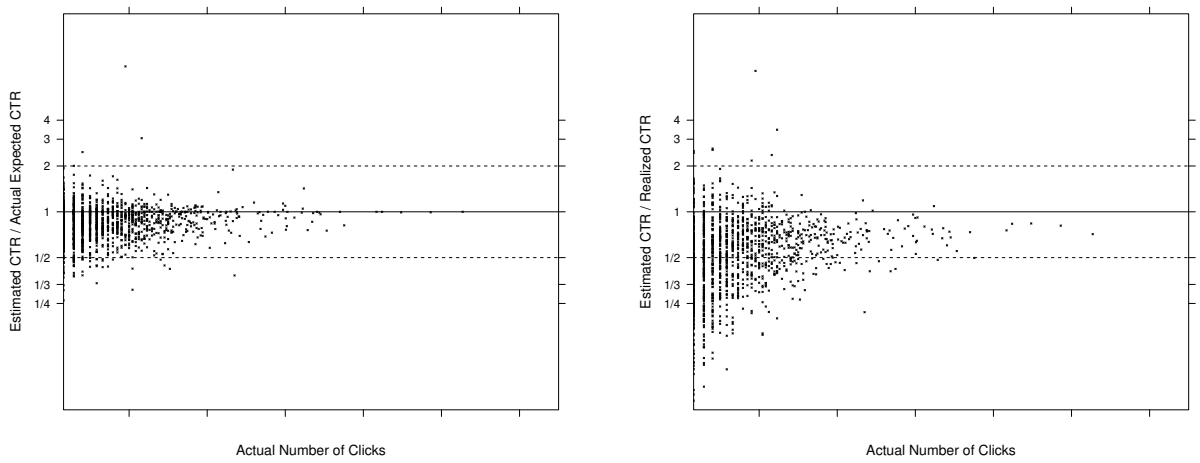
Thus far we have expressed the cost function $c_i(x_i)$ in terms of other stochastic variables (such as distributions over participants and opponent bids), but we have not specified how these other distributions are obtained. We follow the approaches taken by Athey and Nekipelov [1] and Pin and Key [16], in that distributions are created from actual sponsored search records for the given time period. For each sampled account, we get all actual user searches for which a term in the account appeared, and then get all competing advertiser bids for those searches. For each term, histograms are created of opponent weighted bids, advertiser reserve prices, number of opponents and number of available promoted slots, and slot click-through rates. Bids and reserve prices are discretized by ten cents, which is the minimum bid increment on Yahoo.

5.2 Cost Function Evaluation

We now evaluate the accuracy of our cost function, following the same evaluation technique used by Pin and Key [16] for comparison purposes. For each term i in the 100 sampled accounts, we create predictions of expected CTR $\hat{x}_i(b)$ using the method described above, and evaluate these predictions against two metrics, both based on the actual sponsored search data from individual auctions.

First, we compare against the *expected slot CTR*, given the advertiser's historically realized slots for that term. This quantifies the error associated with making the independence assumptions described above and with assuming a fixed bid (taken to be the average) over the entire duration of the month. Put another way, this gives an idea of how accurately an advertiser might be able to predict CTR, given the limitations of not knowing *a priori* the realized values of stochastic processes. Figure 5(a) shows the relative error in estimating CTR. Each term is a single data point, and terms are sorted on the x-axis by the number of actual clicks received for that term. We observe that the mean and median ratios between predicted and actual expected CTR are 0.99 and 0.97, respectively. As observed by Pin and Key, we also find that the variance in relative error across terms decreases as the number of clicks increases.

Next, we compare our estimates to the *realized slot CTR*. This takes into account the fact that our predictions of an advertiser's CTR may not be perfectly accurate, even if we know exactly the slot that the advertiser appears in. Figure 5(b) shows, for each term, the estimated clicks for placing the bid versus the actual number of clicks. We see a bias in these predictions, as our model is under-predicting the amount of clicks that will occur. This can be explained by our account selection process: since accounts with the most clicks were favored in the sampling, these sampled accounts



(a) Predicted versus expected CTR.

(b) Predicted versus realized CTR.

Figure 5: Accuracy of the cost function's CTR estimates when compared to the expected and realized CTRs from actual sponsored search logs. Plots show a random sub-sampling of 10% of all terms for clarity. Model accuracy is shown with respect to the number of clicks received by each term. The x-axis is on a logarithmic scale and labels ommitted for confidentiality reasons.

are more likely to have received more clicks than expected. Advertisers whose terms experience this bias would likely be unable to predict such a deviation from a learned model of their click-through rate, particularly if the increased amount of clicks is due to stochastic user behavior and not any changes made to the advertiser's ads or keywords.

Despite independence assumptions about the distributions of opponent bids and the number of participants, our results support the findings of Pin and Key [16] that their model estimates click-through rates with accuracy comparable to other methods proposed in the literature. While we use their model because of its computational tractability and similar accuracy to more complex methods, any advertiser cost function could similarly be substituted into Equation (6) to regress on advertiser values-per-click.

6. EXPERIMENTAL EVALUATION

We fit a separate model to each account to allow for the fact that the effect of the features used might vary across accounts. We evaluate the predictive performance of our approach via synthetic leave-out estimates: for each account, some fixed proportion of the terms, ad groups, or campaigns is used for training, and we then record the aggregate *regret* of the fitted model's value predictions on the remaining terms used for testing. The regret on a term is given by (6)—observe that this does not include the squared loss (8–9) at the ad group and campaign levels used for model fitting, which is only for regularization. We considered training-testing splits where 50%, 75%, and 90% of the terms, ad groups, or campaigns are randomly selected for training. For each combination of leave-out and split we did 10 runs and averaged the results over the runs. Because some accounts have too few terms to properly develop a value model, we restricted our attention to the 88 accounts in our sample with at least 40 terms. The remaining accounts on which we trained and made predictions consisted of nearly 150 thousand terms.

As a baseline prediction for a term's value, we used the average value of the other terms in its ad group, as predicted by the Pin and Key approach. Specifically, the value on each other term i is given by $v_i' = \nabla c_i(x_i)$, where c_i is the cost function derived as explained in Section 4.1 and x_i is the observed CTR. The regret on the term is again evaluated using (6). If there are no other terms in the given term's ad group or campaign present in the training set (e.g., when leaving out ad groups or campaigns), we use the average value over terms in the same campaign or over the account, respectively.

6.1 Features

We used the following features to develop our models. Features can arise at the term, ad group, and campaign levels, but there were no obvious features to use at the campaign level so only an intercept is present there for regularization. The following predictors (or categories of predictors) were used at the term level. As indicated, some predictors are logged to account for skew in their distributions. No other transformations were made to these predictors and no interaction effects were introduced.

age five separate predictors indicating the percent of offers to users in their twenties, thirties, etc.

gender three separate predictors indicating the percent of offers to male and female users, or gender unknown.

exact match percent of opponent ads presented due to exact rather than advanced match.

user click propensity (log) mean of a metric quantifying, for each keyword the ad matches to, the propensity of users who search on the keyword to click on ads—see [12].

query length mean word length of the user queries leading to the offer.

north state five separate predictors indicating the percent of searches for the keyword yielding a page with zero, one, two, three, or four ads at the top of the page.

competitors (log) mean number of competitors on the keyword.

The first four features capture aspects of a term that may have inherent value to the advertiser (e.g., the demographic profile of searchers), while the last two capture the level of competition on the term, which may correlate with unseen aspects of the term that drive value. Statistics on gender or age come from the self-reports of searchers who are logged in to their Yahoo account.[7] We next turn to the predictors at the ad group level.

creatives number of creatives in the ad group.

linead length word length of the keyword bid on.

title length word length of the ad title.

These predictors give some indication of the specificity of the advertisement, which may correlate with value. To clarify the difference between 'linead length' and 'query length', the former corresponds to the length of the keyword specified for the ad group, while the latter corresponds to the mean length of the query actually matched to, which can vary in the case of advanced match.

In terms of goodness of fit, we found that the most predictive feature varied across accounts. This result is not surprising, as different accounts do not necessarily represent companies from the same industry and thus could have different levels of importance for different user characteristics.

6.2 Results

When evaluating the model and baseline on the testing set terms we recorded the average regret per term for each and compared them against each other. We consider two comparison metrics:

$$\text{absolute improvement} = \text{baseline regret - model regret}$$

$$\text{relative improvement} = \frac{\text{baseline regret - model regret}}{\text{baseline regret}}$$

A positive value for either indicates that the model outperformed the baseline, and vice-versa. Note however that there is an asymmetry in reporting relative improvement: its maximum value is 1 (regret is always non-negative), or 100%, but it can take on arbitrarily large negative values as the baseline regret approaches zero. As a result, taking the mean of the relative improvement across accounts leads to a misleading assessment of model performance: the result is highly negative for every choice of split and choice of leave-out (i.e., term, ad group, or campaign) due to outliers.

[7]Personal account information was not accessed for this study. Anonymized, aggregate statistics on gender and age were already available in the sponsored search logs.

To give a clearer picture of the distribution, Table 2 summarizes for each regime the relative improvement quartiles, which are robust to outliers. Recall that the first quartile cuts off the lowest 25% of the data, the third cuts off the highest 25%, and the second quartile is the median. The interquartile range (difference between third and first quartiles) is a non-parametric analog of standard deviation.

Leave-out	Split	Q_1	Q_2	Q_3	IQR
term	50%	-130.9	-3.9	30.8	161.8
	75%	-103.9	4.1	33.8	137.7
	90%	-121.2	5.5	31.4	152.7
ad group	50%	-107.4	-5.1	29.8	137.2
	75%	-87.5	6.7	36.5	124.1
	90%	-79.3	4.6	40.2	119.6
campaign	50%	-24.6	13.9	38.6	63.2
	75%	-23.1	14.4	43.4	66.5
	90%	-26.9	11.3	45.6	72.5

Table 2: Quartiles of percent relative improvement over the baseline, across accounts, together with interquartile range.

The clearest observation is that the relative performance against the baseline improves as one moves up the leave-out hierarchy from terms to ad groups to campaigns. We attribute this to the fact that model performance remains relatively stable in each case while the baseline necessarily degrades. There is a monotonic improvement in model performance as the size of the training set increases, as expected, but relative improvement is not necessarily monotone because the baseline also improves. We see from the second quartile (median) that the model improves on the baseline for over half the accounts in all regimes except leaving out terms or ad groups with 50% training split.

In Figure 6 we present summaries of the model's absolute improvement over the baseline. Observe that the differences are very small: for many they are just fractions of cents more in profit. To give a sense of the scale, for a 90% split the median improvement was 0.01, 0.14, and 0.23 for terms, ad groups, and campaigns. However, given the click volume of these accounts small differences can translate into substantial increases in monthly profits. In many cases the opportunities for improvement may be limited, for instance when the cost function c_i on a term is almost flat. The same pattern of improvement as earlier can be seen when moving from term to ad group to campaign value prediction, as the right tail of the improvement distribution becomes slightly heavier.

We conclude from this analysis that our modeling approach holds the most promise for predicting advertiser values on newly created ad groups and campaigns, rather than just single terms added to ad groups. In the latter case, the value for the term is likely close to the average value of the ad group, which is expected given the close relationship between terms in an ad group—recall they all share the same creative.

7. CONCLUSIONS

This paper proposed a regret-based hierarchical model for estimating advertiser values per click from keyword char-

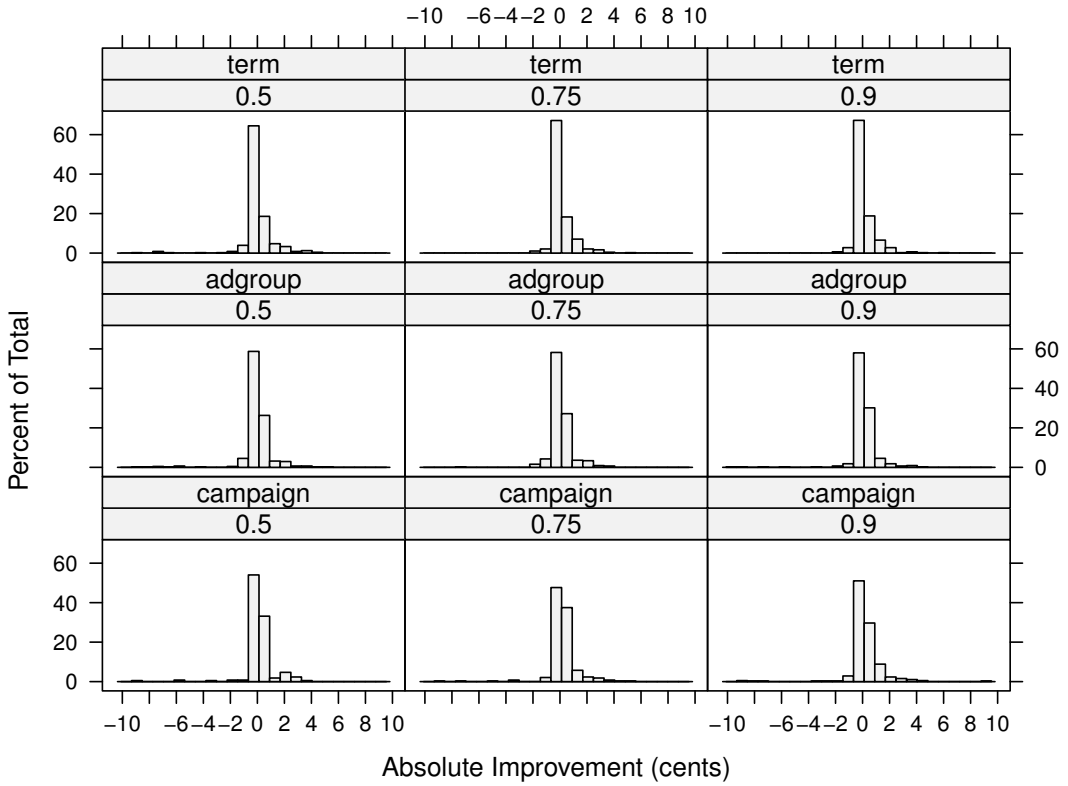


Figure 6: Distribution of the absolute improvement of the model over the baseline. A positive difference indicates the model did better than the baseline.

acteristics, observed costs and click-through rates, and observed advertiser bids. Our modeling strategy was evaluated on nearly 150 thousand terms and outperformed a competitive baseline [16] on the majority of our leave-out experiments. We found that value estimation using this approach is most fruitful when predicting values on new ad groups and campaigns. We also independently validated recently proposed methods for estimating advertiser cost and click beliefs, and provided data-driven insights into the structure of advertisers' sponsored search campaigns.

We see several avenues for improvement and future work. First, there is room for improvement in the prediction performance of our hierarchical model. We chose this kind of model with a view towards interpretation as well as prediction, which can be important if advertisers demand explanations for keyword or bid suggestions. We believe good improvements could be obtained using machine learning algorithms specialized for prediction (e.g, boosting [18]) if that were the sole concern. We also see the need to move beyond ad-hoc feature selection. To this end, we intend to apply techniques such as topic models [4] to uncover conceptual and semantic regularities among campaign terms.

8. ACKNOWLEDGEMENTS

We thank Amy Greenwald and David Pennock for initiating this project, and Eliot Li for bringing the collaborators together. Furcy Pin gave us valuable clarifications on his clicks and cost modeling methodology. We received helpful comments and suggestions from Amy Greenwald, Patrick Jordan, Ashvin Kannan, Prabhakhar Krishnamurthy, Eren Manavoglu, David Pennock, and Michael Schwarz.

References

[1] Susan Athey and Denis Nekipelov. A structural model of sponsored search advertising auctions. Technical report, Microsoft Research, May 2010.

[2] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergence. *Journal of Machine Learning Research*, 6:1–48, 2005.

[3] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. In *Journal of Machine Learning Research*, volume 3, pages 993–1022, March 2003.

[5] Andrei Broder, Evgeniy Gabrilovich, Vanja Josifovski, George Mavromatis, and Alex Smola. Bid generation for advanced match in sponsored search. In

Proceedings of the fourth ACM International Conference on Web Search and Data Mining, pages 515–524, 2011.

[6] Yifan Chen, Gui-Rong Xue, and Yong Yu. Advertising keyword suggestion based on concept hierarchy. In *Proceedings of the International Conference on Web Search and Web Data Mining*, pages 251–260, 2008.

[7] Quang Duong and Sébastien Lahaie. Discrete choice models of bidder behavior in sponsored search. In *Proceedings of the 7th International Workshop on Internet and Network Economics*, 2011.

[8] Benjamin Edelman. Strategic bidder behavior in sponsored search auctions. In *Workshop on Sponsored Search Auctions*, pages 192–198, 2005.

[9] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the Generalized Second-Price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1), March 2007.

[10] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.

[11] Anindya Ghose and Sha Yang. An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, 55: 1605–1622, October 2009.

[12] Dustin Hillard, Stefan Schroedl, Eren Manavoglu, Hema Raghavan, and Chris Leggetter. Improving ad relevance in sponsored search. In *Proceedings of the third ACM international conference on Web Search and Data Mining*, pages 361–370, New York, NY, 2010. ACM.

[13] Bernard J. Jansen and Lauren Solomon. Gender demographic targeting in sponsored search. Working paper, 2010.

[14] Sébastien Lahaie and David M. Pennock. Revenue analysis of a family of ranking rules for keyword auctions. In *Proceedings of the 8th ACM Conference on Electronic Commerce*, pages 50–56, 2007.

[15] Sébastien Lahaie, David M. Pennock, Amin Saberi, and Rakesh V. Vohra. Sponsored search auctions. In Noam Nisan, Tim Roughgarden, Éva Taros, and Vijay V. Vazirani, editors, *Algorithmic Game Theory*, pages 699–716. Cambridge University Press, 2007.

[16] Furcy Pin and Peter Key. Stochastic variability in sponsored search auctions: observations and models. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, pages 61–70, 2011.

[17] Oliver J. Rutz and Randolph E. Bucklin. A model of individual keyword performance in paid search advertising. *SSRN eLibrary*, 2007.

[18] Robert E. Schapire. A brief introduction to boosting. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 1401–1406, 1999.

[19] Hal R. Varian. Goodness-of-fit in optimizing models. *Journal of Econometrics*, 46:125–140, 1990.

[20] Hal R. Varian. Position auctions. *International Journal of Industrial Organization*, 25:1163–1178, 2007.

[21] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th International World Wide Web Conference*, pages 261–270, Madrid, Spain, 2009.